bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Analysis of gene expression and connectivity on hippocampus of Alzheimer's disease by a new comprehensive approach

Shan Jiang[1,*]

[1]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA

*Corresponding author at: Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. E-mail address: shannjiang@hotmail.com (Shan Jiang).

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Gene expression and gene connectivity describe two different functional aspects of a gene. These two different measures reveal different information about the involvement of genes in disorders. Previous case-control gene expression studies have often focused on expression level of individual genes. Correlated expression relationships among genes, measured as gene connectivity, have obtained limited attention. We developed a comprehensive method, TRIple Differentiation (TRID), to assess these two measures, both separately and jointly. We applied TRID to gene expression data in hippocampus tissue samples from three Alzheimer's disease (AD) microarray datasets. Following TRID, comparisons among the three datasets showed poor consistency for disease-associated individual genes but reproducible changes of disease-associated biological pathways annotated for functional protein-protein interaction (PPI) modules identified from network analysis. Our results suggest that changes of gene expression in hippocampus of AD patients are highly heterogeneous at the individual gene level, while biological pathways annotated for PPI modules identified based on TRID weights demonstrate consistency among the three datasets.

Key words: gene expression; gene connectivity; Alzheimer's disease; protein-protein interaction network; functional module.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

1. Introduction

Investigating change of gene expression level (inner trait of a gene) in patients in contrast to healthy controls is the conventional approach to identify disease-associated gene. However, differential expression does not capture change with respect to gene relationship (inter trait of a gene) between cases and controls. To study change of gene relationship, Horvath and colleagues developed Weighted Gene Co-expression Network Analysis (WGCNA) a decade ago. In common practice, WGCNA identifies clusters of co-expressed genes based on correlated expression. Differential expression between cases and controls is then evaluated for eigengene of each module, which represent the expression level of an entire module (Langfelder and Horvath, 2008; Zhang and Horvath, 2005). Therefore, WGCNA still examines change of expression level in nature; the concept of gene relationship only applied to the construction of co-expression network.

To supplement WGCNA, Fuller *et al*. introduced the concept and calculation of gene connectivity (Fuller et al., 2007). Gene connectivity was defined as the summed co-expression levels of a specific gene with all other genes (Fuller et al., 2007). Gene connectivity can be calculated in cases and controls separately and then compared to assess differential connectivity of individual genes (Fuller et al., 2007). Differential connectivity has been applied successfully (Doering et al., 2012; Fuller et al., 2007).

However, analyzing either differential expression or differential connectivity alone may miss genes whose changes are not significant uni-dimensionally but stand out if and when assessed jointly. It is desirable to investigate disturbance of genes with regards to differential expression and differential connectivity simultaneously to avoid the missed opportunity.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this study, we developed a new method TRIple Differentiation (TRID), which includes three analyses: differential expression, differential connectivity and a fusion analysis of both differential expression and differential connectivity. The fusion analysis was termed Differential COnnectivity and Differential Expression (DiCODE). We comprehensively assessed changes of genes in hippocampus of Alzheimer's disease (AD) patients and identified highly consistent biological processes or pathways among multiple datasets by TRID.

2. Methods

2.1. AD hippocampus expression data collection and quality control

Microarray expression data regarding hippocampus of AD patients were retrieved from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) by searching with the keywords "Alzheimer" and "hippocampus". To reduce uncontrollable bias, only microarray datasets with consistent assay platform were included. After manual inspection and filtration, three datasets containing hippocampus expression data, GSE5281 (10 AD cases and 13 controls), GSE28146 (22 AD cases and 8 controls) and GSE48350 (19 AD cases and 43 controls), were retained for analysis.

Robust multichip average (RMA) normalization was performed on the three microarray datasets using R package affy (version 1.57.0).

2.2. Gene weighting strategies

To apply TRID to the three datasets, we began with the classical differential expression analysis between cases and controls, which was termed as the differential expression test in TRID. Welch's t-test was used to assess differential expression levels of individual genes between cases and controls (Doering et al., 2012). Absolute value of

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the t-statistic, denoted by $|T(i)|$, was used to represent the differential expression (DE) weight of the *i*th gene, where a higher absolute t-statistic value corresponds to a greater differential expression level.

For the differential connectivity (DC) test in TRID, connectivity levels of individual genes in cases and controls were calculated respectively by using $k_{(case)}(i) = \sum_{j=1}^{n}|p_{ij}|$ and $k_{(control)}(i) = \sum_{j=1}^{n}|p_{ij}|$, where the p_{ij} denotes the Pearson correlation coefficient between the *i*th and *j*th genes. To facilitate the comparison of connectivity level of the *i*th gene between cases and controls, connectivity level of the *i*th gene in cases and controls were standardized through z-score transformation such that $K_{case}(i) = \frac{k_{case}(i)-mean(k_{case})}{\sigma(k_{case})}$ and $K_{control}(i) = \frac{k_{control}(i)-mean(k_{control})}{\sigma(k_{control})}$. $DiffK(i) = K_{case}(i) - K_{control}(i)$ was defined as differential connectivity of the *i*th gene and $|DiffK(i)|$ as the differential connectivity weight of the *i*th gene, as defined by Fuller *et al* (Fuller et al., 2007).

For the combined Differential COnnectivity and Differential Expression (DiCODE) test in TRID, in order to fuse differential expression and differential connectivity weights calculated for the *i*th gene, the differential expression weight $|T(i)|$ and differential connectivity weight $|DiffK(i)|$ were standardized through z-score transformation such that $T_{scaled}(i) = \frac{|T(i)|-mean[|T(i)|]}{\sigma[|T(i)|]}$ and $DiffK_{scaled}(i) = \frac{|DiffK(i)|-mean[|DiffK(i)|]}{\sigma[|DiffK(i)|]}$. $DiCODE(i) = T_{scaled}(i) + DiffK_{scaled}(i)$ was defined as the DiCODE weight of the *i*th gene.

We then extracted top 5% high ranking DE, DC and DiCODE weighted genes of the three datasets respectively. For each weighting strategy, venn diagram was used to

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

show the shared and specific top 5% high ranking genes among the three datasets.

The R package TRID can be downloaded from GitHub (https://github.com/shannjiang/TRID).

2.3. TRID weight-based functional PPI module identification and gene ontology (GO) annotation

DE, DC and DiCODE weights were termed TRID weights in this study. Cytoscape plugin jActiveModules was used to identify significantly enriched functional PPI modules based on TRID weights of genes (Ideker et al., 2002). In this study, we used PPI information from the IntAct and MINT databases, which detail molecular interactions in well-studied organisms such as *S. cerevisiae* and *Homo sapiens*, to construct a combined background PPI network (Cline et al., 2007; Hermjakob et al., 2004; Zanzoni et al., 2002).

DAVID was then used to annotate the functions of identified PPI modules (Huang et al., 2009).

3. Results

3.1. TRID weights of each gene generated from hippocampus microarray datasets

For TRID weights of individual genes generated from GSE5281, GSE28146 and GSE48350 microarray datasets, please refer to supplementary Table 1, 2 and 3 respectively.

For DE, DC and DiCODE weighting strategies, venn diagrams of shared and specific top 5% high ranking weighted genes among the three datasets were shown in Fig. 1. For DE weighting strategy, 19 genes were shared among the three datasets (Fig. 1A). For both DC and DiCODE weighting strategy, only 3 genes were shared among the three

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

datasets (Fig. 1B and 1C).

3.2. TRID weights-based functional PPI module identification and GO annotation

TRID weights-based functional PPI modules were identified for individual datasets. For the network graphs of functional PPI modules of individual datasets identified based on DE, DC and DiCODE weights, please refer to Supplementary Figures 1, 2, and 3 respectively. For detailed GO annotations of these identified functional PPI modules of individual datasets based on DE, DC and DiCODE weights, please refer to Supplementary Table 4, 5 and 6 respectively. Venn diagrams of shared and specific significantly enriched GO terms among the three datasets based on DE, DC and DiCODE weights were shown in Fig. 2A, 2B and 2C respectively. Nearly 30 to 50% of the significantly enriched GO terms were shared among the three datasets based on each weight of TRID. For detailed significant GO terms shared among the three datasets please refer to Supplementary Tables 7, 8 and 9 for DE, DC and DiCODE weighting strategies respectively. Cell nucleus-associated biological processes, cell components and molecular functions were highly shared for individual weighting strategies of TRID. Nuclear lamins are fibrous proteins providing structural function and transcriptional regulation in the cell nucleus (Dechat et al., 2008). Dysfunctions of and mutations in the genes encoding A or B-type lamins, or termed laminopathy, were associated to neurodegeneration recently (Frost, 2016; Frost et al., 2016).

4. Discussion

In this study we developed a new method called TRID, which includes differential expression analysis, differential connectivity analysis and DiCODE analysis for individual genes. With DiCODE analysis, we can detect genes with changes that

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

could be missed in conventional differential expression analysis, emphasizing the combined effects instead of either one alone.

Only less than 2% of top 5% high TRID-weighted genes were shared among the three datasets (Fig. 1), which suggests heterogeneity among samples. More than 90% of disease-associated SNPs are located in non-coding regions of the genome (Hindorff et al., 2009), suggesting the potential regulatory nature of these SNPs. The International Genomics of Alzheimer's Project (IGAP) GWAS meta-analysis identified 20 loci associated with AD (Lambert et al., 2013). In this study, most of the AD hits lie outside of gene coding regions. At least some of them may contribute to changes in gene expression (Knight, 2003). Variable combinations of risk alleles at different non-coding SNP loci in patients contribute to expression profile heterogeneity in AD. Moreover, environment, including social and psychological factors (Read et al., 2004), can potentially influence gene expression. Some research have implied that epigenetic modifications largely influence regulations of genes in AD (Kwok, 2010; Rao et al., 2012), which could further contribute to AD heterogeneity or individual differences of expression profiles.

In contrast to the high heterogeneity at individual gene level, we observed very robust findings of biological processes or signaling pathways from functional PPI module identification and annotation, even though all datasets used in this study had relatively small sample sizes. It indicates that different genes responsible for the underlying etiology of AD involve in the same pathways or PPI modules. Methodologically, by coupling to TRID weighting, functional PPI module identification and annotation could be an effective way to uncover pathways involved in AD progression.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In summary, our study revealed that (1) TRID can weight genes effectively and comprehensively; (2) coupling PPI module analyses with TRID weighting may help to detect the converged changes in AD; and (3) AD is heterogeneous at individual gene level, but many differential genes are involved in the same nucleus-associated pathways and disease regulatory modules.

Conflict of interest

The authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B., Hanspers, K., Isserlin, R., Kelley, R., Killcoyne, S., Lotia, S., Maere, S., Morris, J., Ono, K., Pavlovic, V., Pico, A. R., Vailaya, A., Wang, P. L., Adler, A., Conklin, B. R., Hood, L., Kuiper, M., Sander, C., Schmulevich, I., Schwikowski, B., Warner, G. J., Ideker, T., Bader, G. D., 2007. Integration of biological networks and gene expression data using Cytoscape. Nat. Protoc. 2, 2366-2382.

Dechat, T., Pfleghaar, K., Sengupta, K., Shimi, T., Shumaker, D. K., Solimando, L., Goldman, R. D., 2008. Nuclear lamins: major factors in the structural organization and function of the nucleus and chromatin. Genes Dev. 22, 832-853.

Doering, T. A., Crawford, A., Angelosanto, J. M., Paley, M. A., Ziegler, C. G., Wherry, E. J., 2012. Network analysis reveals centrally connected genes and pathways involved in CD8+ T cell exhaustion versus memory. Immunity 37, 1130-1144.

Frost, B., 2016. Alzheimer's disease: An acquired neurodegenerative laminopathy. Nucleus 7, 275-283.

Frost, B., Bardai, F. H., Feany, M. B., 2016. Lamin Dysfunction Mediates Neurodegeneration in Tauopathies. Curr. Biol. 26, 129-136.

Fuller, T. F., Ghazalpour, A., Aten, J. E., Drake, T. A., Lusis, A. J., Horvath, S., 2007. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm. Genome 18, 463-472.

Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., Apweiler, R., 2004. IntAct: an open source molecular interaction database. Nucleic Acids Res. 32, D452-455.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., Manolio, T. A., 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. 106, 9362-9367.

Huang, D. W., Sherman, B. T., Lempicki, R. A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. 4, 44-57.

Ideker, T., Ozier, O., Schwikowski, B., Siegel, A. F., 2002. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics 18 Suppl 1, S233-240.

Knight, J. C., 2003. Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. Clin. Sci. 104, 493-501.

Kwok, J. B., 2010. Role of epigenetics in Alzheimer's and Parkinson's disease. Epigenomics 2, 671-682.

Lambert, J. C., Ibrahim-Verbaas, C. A., Harold, D., Naj, A. C., Sims, R., Bellenguez, C., DeStafano, A. L., Bis, J. C., Beecham, G. W., Grenier-Boley, B., Russo, G., Thorton-Wells, T. A., Jones, N., Smith, A. V., Chouraki, V., Thomas, C., Ikram, M. A., Zelenika, D., Vardarajan, B. N., Kamatani, Y., Lin, C. F., Gerrish, A., Schmidt, H., Kunkle, B., Dunstan, M. L., Ruiz, A., Bihoreau, M. T., Choi, S. H., Reitz, C., Pasquier, F., Cruchaga, C., Craig, D., Amin, N., Berr, C., Lopez, O. L., De Jager, P. L., Deramecourt, V., Johnston, J. A., Evans, D., Lovestone, S.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Letenneur, L., Moron, F. J., Rubinsztein, D. C., Eiriksdottir, G., Sleegers, K., Goate, A. M., Fievet, N., Huentelman, M. W., Gill, M., Brown, K., Kamboh, M. I., Keller, L., Barberger-Gateau, P., McGuiness, B., Larson, E. B., Green, R., Myers, A. J., Dufouil, C., Todd, S., Wallon, D., Love, S., Rogaeva, E., Gallacher, J., St George-Hyslop, P., Clarimon, J., Lleo, A., Bayer, A., Tsuang, D. W., Yu, L., Tsolaki, M., Bossu, P., Spalletta, G., Proitsi, P., Collinge, J., Sorbi, S., Sanchez-Garcia, F., Fox, N. C., Hardy, J., Deniz Naranjo, M. C., Bosco, P., Clarke, R., Brayne, C., Galimberti, D., Mancuso, M., Matthews, F., European Alzheimer's Disease, I., Genetic, Environmental Risk in Alzheimer's, D., Alzheimer's Disease Genetic, C., Cohorts for, H., Aging Research in Genomic, E., Moebus, S., Mecocci, P., Del Zompo, M., Maier, W., Hampel, H., Pilotto, A., Bullido, M., Panza, F., Caffarra, P., Nacmias, B., Gilbert, J. R., Mayhaus, M., Lannefelt, L., Hakonarson, H., Pichler, S., Carrasquillo, M. M., Ingelsson, M., Beekly, D., Alvarez, V., Zou, F., Valladares, O., Younkin, S. G., Coto, E., Hamilton-Nelson, K. L., Gu, W., Razquin, C., Pastor, P., Mateo, I., Owen, M. J., Faber, K. M., Jonsson, P. V., Combarros, O., O'Donovan, M. C., Cantwell, L. B., Soininen, H., Blacker, D., Mead, S., Mosley, T. H., Jr., Bennett, D. A., Harris, T. B., Fratiglioni, L., Holmes, C., de Bruijn, R. F., Passmore, P., Montine, T. J., Bettens, K., Rotter, J. I., Brice, A., Morgan, K., Foroud, T. M., Kukull, W. A., Hannequin, D., Powell, J. F., Nalls, M. A., Ritchie, K., Lunetta, K. L., Kauwe, J. S., Boerwinkle, E., Riemenschneider, M., Boada, M., Hiltuenen, M., Martin, E. R., Schmidt, R., Rujescu, D., Wang, L. S., Dartigues, J. F., Mayeux, R., Tzourio, C., Hofman, A., Nothen, M. M., Graff, C., Psaty, B. M., Jones, L., Haines, J. L., Holmans, P. A.,

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Lathrop, M., Pericak-Vance, M. A., Launer, L. J., Farrer, L. A., van Duijn, C. M., Van Broeckhoven, C., Moskvina, V., Seshadri, S., Williams, J., Schellenberg, G. D., Amouyel, P., 2013. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. Nat. Genet. 45, 1452-1458.

Langfelder, P., Horvath, S., 2008. WGCNA: an R package for weighted correlation network analysis. BMC Bioinform. 9, 559.

Rao, J. S., Keleshian, V. L., Klein, S., Rapoport, S. I., 2012. Epigenetic modifications in frontal cortex from Alzheimer's disease and bipolar disorder patients. Transl. Psychiatry 2, e132.

Read, J., Mosher, L. R., Bentall, R. P. Models of madness: Psychological, social and biological approaches to schizophrenia. 1st ed. London: Routledge; 2004.

Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., Cesareni, G., 2002. MINT: a Molecular INTeraction database. FEBS Lett. 513, 135-140.

Zhang, B., Horvath, S., 2005. A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4, Article17.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

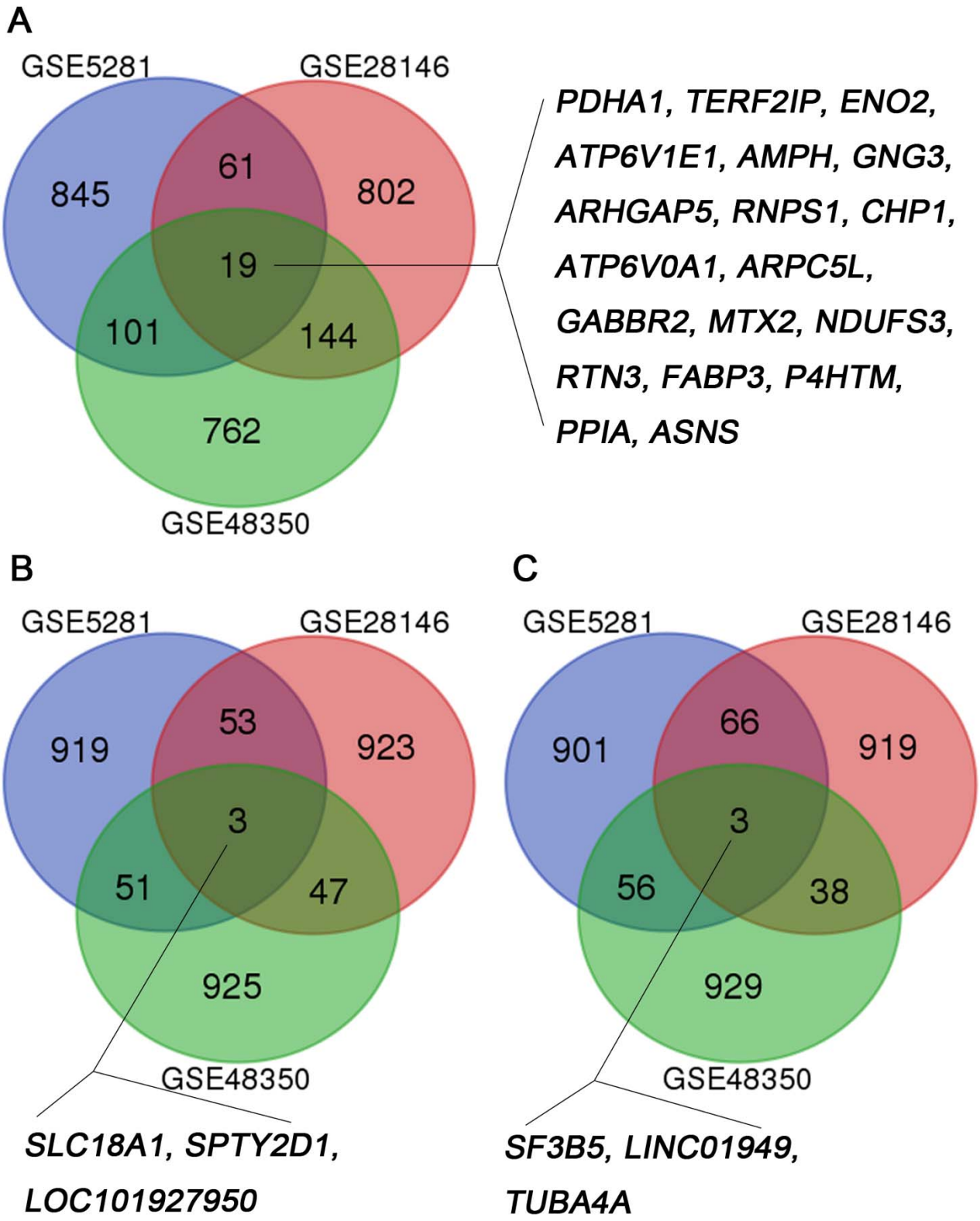
Figure Legends

Fig. 1. Venn diagrams of shared and specific top 5% high ranking weighted genes among the three hippocampus microarray datasets based on TRID DE (A), DC (B) and DiCODE (C) weighting strategies.

Fig. 2. Venn diagrams of shared and specific GO term annotations for identified functional PPI modules based on TRID DE (A), DC (B) and DiCODE (C) weights.

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1



bioRxiv preprint doi: https://doi.org/10.1101/2020.01.14.906446; this version posted January 15, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

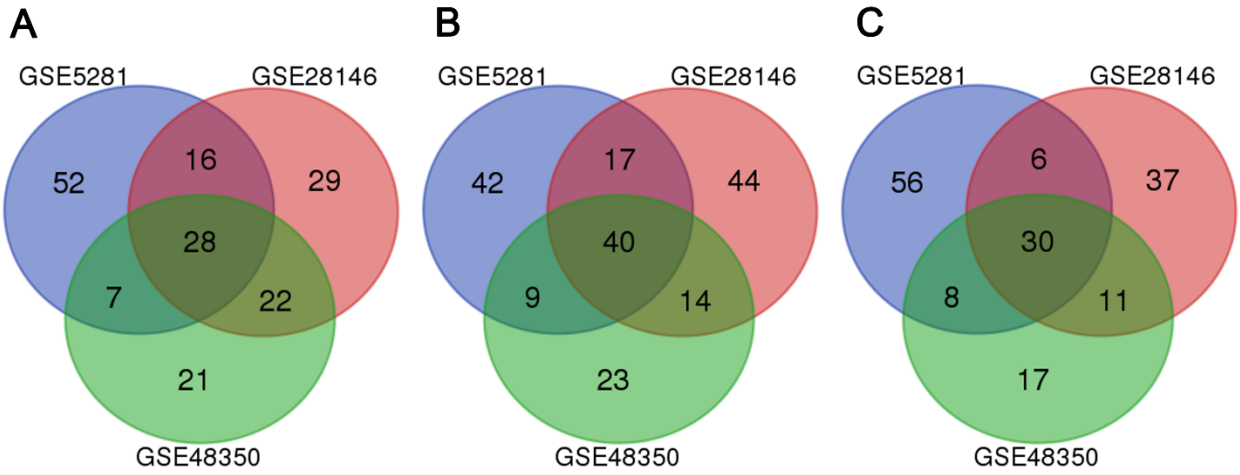



Figure 2